Emergence du langage par jeux déictiques dans une société d'agents sensori-moteurs en interaction

Clément Moulin-Frier, Jean-Luc Schwartz, Julien Diard, Pierre Bessière

Emergence du langage par jeux déictiques dans une société d'agents sensori-moteurs en interaction

Clément Moulin-Frier †, *Jean-Luc Schwartz* †, *Julien Diard* ††, *Pierre Bessière* †††

CNRS - Université de Grenoble

† GIPSA-Lab, ICP, UMR 5216 Prenom.Nom@gipsa-lab.inpg.fr

†† LPNC, UMR 5105 Julien.Diard@upmf-grenoble.fr

††† LIG-Lab, UMR 5217 Pierre.Bessiere@imag.fr

ABSTRACT

In this paper, we show how some properties of human language could emerge from the primitive deixis function. For this aim, we model a society of sensori-motor agents able to produce vocalizations and to point to objects in their environnement. We show how principles of the Dispersion Theory [6] and the Quantal Theory [13] could emerge from the interaction between these agents.

Keywords: Evolutionnary Linguistic, Language Emergence, Bayesian Modeling, Cognitive Robotics, Multi-Agents Systems

1. Introduction

Depuis les années 70, des linguistes s'interrogent sur la possibilité de dériver le langage du non-langage [6]. Ils ouvrent ainsi la voie à des théories du langage "orientées-substance", qui mettent en valeur le rôle des contraintes morphologiques et cognitives dans l'apparition de formes universelles dans les langues du monde.

De cette évolution scientifique est né récemment un nouveau champ de recherche, parfois nommé Linguistique Evolutionnaire, qui cherche à montrer par la simulation informatique comment ces formes universelles peuvent émerger dans une société d'agents virtuels [12, 2]. C'est de cette problématique que traite cet article, en se basant sur des hypothèses fortes concernant l'apparition du langage chez l'Homme issues de certaines théories orientées-substance.

En effet, dans cet ensemble de théories, nous pouvons distinguer deux types d'approches. Le premier consiste à faire l'observation que le langage existe, puis à tenter d'expliquer ses tendances universelles par des contraintes périphériques au langage, provenant typiquement des propriétés des systèmes de perception et de production de la parole, sous l'hypothèse que les propriétés du langage émergent des conditions de communication. C'est le point de départ de théories telles que la théorie de la dispersion de Lindblom [6] ou la théorie quantique de Stevens [13]. Le deuxième type d'approches consiste quant à lui à tenter d'expliquer le langage lui-même par des fonctions primitives dont il dériverait, telles la mastication (théorie Frame/Content [7]) ou la déixis [1] (théorie Vocalize to Localize [1]). Nous les nommerons respectivement "théories de la morphogénèse" et "théories de l'origine".

Notre hypothèse centrale est que les théories de la morphogénèse doivent elles-même dériver des théories de l'origine. Nous nous situons dans le cadre "Vocalize to Localize" [1] dans lequel la déixis a constitué un point de départ, un bootstrap à la naissance d'un langage articulé. Dans ce cadre, nous modélisons une société d'agents probabilistes (bayésiens) capables de montrer des objets dans un environnement. Nous montrons alors comment un code de parole respectant certains principes de la théorie de la dispersion et de la théorie quantique peut émerger de cette fonction de déixis.

2. Paradigme d'interaction

L'introduction d'une fonction de déixis dans nos agents est l'une des principales particularités de notre approche. En effet, nous pensons que le langage est né du besoin de "montrer de la voix". C'est la théorie "Vocalize to Localize" [1] selon laquelle la parole se serait constituée autour de trois composantes fondamentales :

- Un système de production d'actions orofaciales pour émettre des vocalisations ;
- Un système de perception phonétique pour structurer le flux sonore en unités perceptives intelligibles ;
- Un système de pointage et d'attention partagée pour montrer le monde et produire du sens.

C'est ce que nous modélisons en plongeant des agents capables de produire des gestes moteurs et de percevoir les percepts correspondants dans un environnement peuplé d'objets à nommer. Des agents sensori-moteurs évoluent donc dans un environnement dans lequel se trouvent des objets qu'ils peuvent identifier. Au cours du temps, ils se retrouvent de façon aléatoire deux par deux devant un objet O. Ils procèdent alors à ce que nous appelons un "jeu déictique", où un des deux agents prend le statut de locuteur, l'autre celui d'auditeur. Pour "montrer de la voix" cet objet, le locuteur propose, par un geste moteur M, une vocalisation. Par les lois de l'acoustique, ce geste moteur produit un percept sensoriel S perçu par l'auditeur (Figure 1). Les deux agents mettent alors à jour leurs connaissances en fonction de cette interaction (c'est-à-dire en fonction des valeurs de O et M pour le locuteur, et de O et S pour l'auditeur). Les jeux déictiques se succèdent au cours du temps, les agents se retrouvant aléatoirement en statut de locuteur ou

[1] Déixis, du grec deiktikos, signifie "action de montrer"

d'auditeur.

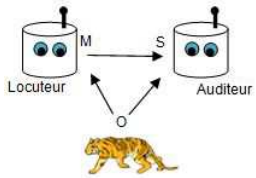

Fig. 1: Un jeu déictique entre deux agents

3. Modélisation

Pour modéliser nos agents, nous utilisons le paradigme de Programmation Bayésienne des Robots (PBR) [5] qui nous permet d'une part d'exprimer nos hypothèses et d'analyser nos résultats de façon mathématiquement claire grâce à des distributions de probabilité, d'autre part d'opérer sur les connaissances ainsi codées grâce à l'inférence bayésienne.

3.1. Fondements mathématiques

La Programmation Bayésienne des Robots vise à spécifier les comportements d'agents sensori-moteurs dans le cadre de la théorie bayésienne des probabilités. Un agent opère sur un ensemble de variables V, correspondant en général à des variables sensorielles (entrées de capteurs) et motrices (valeurs de commande). La méthode PBR se déroule alors en deux phases. La première est une phase déclarative dans laquelle on énonce des connaissances pertinentes au domaine. Elle a pour objectif le calcul de la distribution de probabilité conjointe $P(V)$. La seconde est une phase procédurale dans laquelle l'agent calcule des termes mathématiques d'intérêt en exploitant les connaissances énoncées préalablement. Plus précisément, il utilise la distribution conjointe pour répondre à des questions probabilistes de type $P(X|Y)$ (par exemple, connaissant la valeur de mes variables perceptives, quelle est la distribution de probabilité sur mes variables motrices ?). Dans le cadre de cette article, nous ne pouvons détailler plus précisément notre cadre computationnel, le lecteur intéressé se reportera à [5] pour les aspects généraux de la PBR et à [8] pour un exposé plus détaillé de nos travaux.

3.2. Modélisation des agents

Dans le cadre du paradigme d'interaction exposé Figure 1, nous retenons quatre variables par agent :
- O_L : l'ensemble des objets devant lesquels l'agent s'est retrouvé en situation de locuteur,
- M : l'ensemble des gestes moteurs qu'un agent est capable de produire en situation de locuteur,
- S : l'ensemble des percepts sensoriels qu'un agent peut percevoir en situation d'auditeur,
- O_A : l'ensemble des objets devant lesquels l'agent s'est retrouvé en situation d'auditeur.

Pour chacun des agents, l'objectif est alors de calculer la distribution conjointe sur ses quatre variables : $P(O_L \wedge M \wedge S \wedge O_A)$. Pour cela, nous faisons les hypothèses suivantes :
- H_1 : la présence des objets dans l'environnement est uniforme (tant en statut de locuteur que d'auditeur).
- H_2 : les agents sont capables d'estimer correctement le percept correspondant à un geste donné, ils connaissent donc la distribution $P(S|M)$. Il s'agit d'un modèle direct de copie d'efférence (estimation des paramètres sensoriels à partir des commandes motrices) classique en Sciences Cognitives [3].
- H_3 : les agents apprennent au cours des jeux déictiques la distribution $P(M|O_L)$, c'est-à-dire la distribution des gestes moteurs pour chaque objet devant lequel ils se sont retrouvés en situation de locuteur.
- H_4 : de même, les agents apprennent au cours des jeux déictiques la distribution $P(S|O_A)$, c'est-à-dire la distribution des percepts entendus pour chaque objet devant lequel il se sont retrouvés en situation d'auditeur.

Pour calculer la distribution conjointe, on la décompose en un produit de termes plus simples à spécifier grâce au théorème de Bayes

$$\begin{aligned} &P(O_L \wedge M \wedge S \wedge O_A) \\ &= P(O_L).P(M|O_L).P(S|M \wedge O_L).P(O_A|S \wedge M \wedge O_L) \end{aligned} . \tag{1}$$

puis on la simplifie grâce aux hypothèses ci-dessus. Ainsi :
- H_1 : le terme $P(O_L)$ est constant (il n'interviendra donc pas dans les calculs).
- H_3 : le terme $P(M|O_L)$ est appris par les agents au cours des jeux déictiques. Nous le considérerons comme une famille de distributions gaussiennes (une par valeur de O_L), dont les paramètres à apprendre par l'agent sont les moyennes μ_{O_L} et les variances V_{O_L}. Nous appellerons cette distribution le modèle locuteur.
- H_2 : le terme $P(S|M \wedge O_L)$ se simplifie en $P(S|M)$ et est connu par les agents. Nous considérons qu'il existe une fonction *percept* telle que $S = percept(M)$. La distribution $P(S|M)$ s'exprime alors comme une distribution Dirac, c'est-à-dire $P(S|M) = \begin{cases} 1 & \text{si } S = percept(M), \\ 0 & \text{sinon.} \end{cases}$. Nous appellerons cette distribution le modèle "copie d'efférence".
- H_4 : la distribution $P(O_A|S \wedge M \wedge O_L)$ se simplifie en $P(O_A|S)$. Par application des règles de Bayes, et comme $P(O_A)$ est uniforme (hypothèse H_1), on a

$$P(O_A|S) = \frac{P(S|O_A)}{\sum_{O_A} P(S|O_A)} . \tag{2}$$

où le terme $P(S|O_A)$ est appris par les agents au cours des jeux déictiques. Nous le considérerons comme une famille de distributions gaussiennes (une par valeur de O_A), dont les paramètres à apprendre par l'agent sont les moyennes μ_{O_A} et les variances V_{O_A}. Nous appellerons cette distribution le modèle auditeur.

Nous obtenons alors l'expression simplifiée :

$$P(O_L \wedge M \wedge S \wedge O_A) \propto P(M|O_L).P(S|M).\frac{P(S|O_A)}{\sum_{O_A} P(S|O_A)} . \tag{3}$$

Par cette équation, chaque agent est capable de calculer la distribution de probabilité conjointe sur l'ensemble de ses variables d'intérêt (O_L, M, S et O_A). Ainsi, il est en mesure de calculer n'importe quelle distribution conditionnelle. Remarquons que comme les distributions $P(M|O_L)$ et $P(S|O_A)$ sont apprises au cours des jeux déictiques par chacun des agents,

la distribution conjointe de l'équation (3) évolue au cours du temps. La section suivante montre comment cette évolution peut mener à la création d'un code de parole commun en accord avec certains principes des théories de la morphogénèse.

4. Simulations et analyses

Dans cette section, nous exposons trois simulations correspondant à trois comportements différents pour nos agents. Nous ne nous attachons pas pour l'instant à la modélisation d'un conduit vocal et d'une oreille réaliste. Les variables Met S sont définies sur un espace unidimensionnel borné et la fonction *percept* définissant la transformation articulatori-acoustique est la fonction identité (on a donc $S = M$). Les trois comportements diffèrent alors par la question probabiliste que chaque agent pose à sa distribution conjointe (équation (3)) pour tirer un geste moteur M en situation de locuteur dans un jeu déictique.

4.1. Le comportement réflexe

Ce premier comportement est un comportement naïf qui permet de montrer que la seule prise en compte des intérêts du locuteur ne permet pas de faire émerger un code de parole commun dans la société d'agents. Devant un objet o_i, il consiste simplement en un tirage du geste moteur M selon la distribution $P(M|O_L = o_i)$ (modèle locuteur). Ce comportement favorise donc les gestes déjà souvent tirés devant un objet donné sans se soucier du besoin de l'auditeur. Il ne peut donc mener à l'émergence d'un code de parole commun, comme nous l'observons Figure 2.

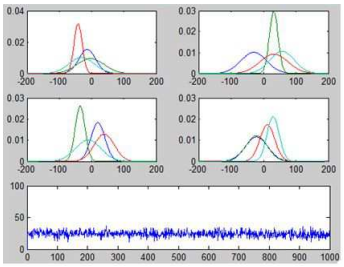

Fig. 2: Comportement réflexe : état de la simulation après 100 000 jeux déictiques dans une société de quatre agents et un environnement de quatre objets. Les quatre fenêtres supérieures correspondent à l'état des quatre agents. Dans chaque fenêtre, les quatre courbes représentent les distributions $P(M|O_L)$ pour chacun des objets de l'environnement. Nous constatons ici qu'il n'y a pas de cohérence dans les gestes produits par chaque agent pour chacun des objets. La fenêtre inférieure représente le taux de compréhension dans la société au cours du temps (avec un facteur 100, 1000=100 000 jeux déictiques), c'est-à-dire le pourcentage de jeux déictiques dans lesquels le percept produit par l'agent locuteur a pu permettre à l'agent auditeur d'inférer le bon objet par la question probabiliste $P(O_A|S)$ (sur les 100 derniers jeux). On observe qu'il reste autour de 25%, ce qui correspond au taux du hasard (il y a quatre objets).

4.2. Le comportement communicatif

Le principe de ce comportement est de tirer un geste moteur qui maximise la probabilité d'être compris par l'auditeur. Ainsi, devant un objet o_i, l'agent locuteur sélectionne un geste M de façon à maximiser la probabilité $P(O_A = o_i|M)$. Or, par inférence bayésienne sur l'équation (3) et d'après la définition de $P(S|M)$, nous avons

$$P(O_A = o_i|M) = P(O_A = o_i|S = percept(M)) \ . \tag{4}$$

Ce comportement va donc chercher à satisfaire le modèle locuteur, en tirant des gestes M dont le percept correspondant permettra à l'auditeur d'inférer facilement l'objet o_i par la question probabiliste $P(O_A|S)$. On observe alors l'émergence d'un code de parole commun (Figure 3).

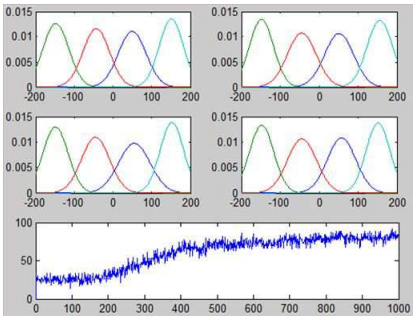

Fig. 3: Comportement communicatif : état de la simulation après 100 000 jeux déictiques. On observe que les agents ont organisé les gestes moteurs produits pour chaque objet de façon satisfaisante (taux de compréhension autour de 80%).

Ce comportement peut être analysé en regard de deux théories de la morphogénèse : la théorie de la dispersion [6] et la théorie quantique [13]. La première stipule que les systèmes phonologiques des langues du monde (en particulier les systèmes de voyelles) tentent de maximiser les distances inter-phonèmes. Nous voyons Figure 3 que cet effet émerge de nos simulations (les $P(M|O_L)$ se sont dispersés, donc les $P(S|O_A)$ également car ici la fonction *percept* est la fonction identité). La théorie quantique quant à elle stipule que les langues du mondes exploitent les non-linéarités présentes dans la transformation articulatori-acoustique en choisissant des phonèmes dans des zones stables, où des variations sur les paramètres moteurs ont peu de conséquences sur les paramètres perceptifs et en utilisant la zone de non-linéarité comme frontière entre catégories phonétiques. Or, si nous introduisons une non-linéarité dans la fonction *percept* en la définissant comme une sigmoïde, nous observons que nos agents choisissent des commandes motrices de part et d'autre de celle-ci (Figure 4).

4.3. Le comportement hybride

Alors que le comportement réflexe cherche seulement à satisfaire le modèle locuteur $P(M|O_L)$ et le comportement communicatif seulement le modèle auditeur $P(O_A|S)$, ce comportement hybride vise à satisfaire

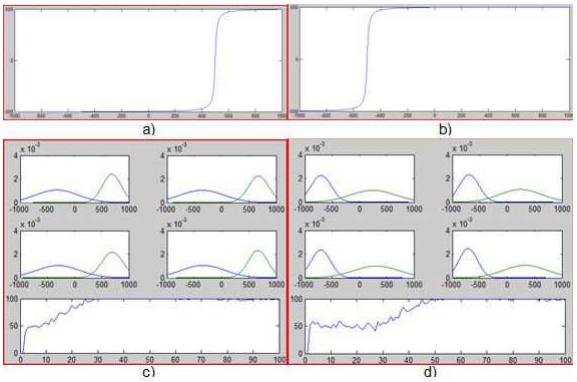

Fig. 4: a) et b) : deux fonctions *percept* avec une non-linéarité (M en abscisse, S en ordonnée). c) et d) : état de la simulation dans une société de quatre agents et un environnement de deux objets après 100 000 jeux déictiques pour chacune des fonctions *percept* a) et b), respectivement. On observe que les agents choisissent des gestes moteurs de part et d'autre de la non-linéarité, et que le taux de compréhension atteint 100%.

les deux, combinant ainsi un comportement conservatif (réflexe) et un comportement dispersif (communicatif). Pour cela, les agents en situation de locuteur devant un objet o_i, tirent un geste moteur M selon la distribution $P(M|O_L = o_i \wedge O_A = o_i)$, c'est-à-dire sélectionne un geste qui à la fois tienne compte de l'objet o_i considéré par le locuteur, et de sa volonté que l'auditeur reconnaisse cet objet. Par inférence bayésienne, cette question probabiliste posée au modèle décrit dans l'équation (3) nous donne :

$$\begin{aligned} &P(M|O_L = o_i \wedge O_A = o_i) \\ &= P(M|O_L = o_i).P(O_A = o_i|S = percept(M)) \end{aligned} \quad . \tag{5}$$

Cette distribution est donc le produit de deux termes, le premier $P(M|O_L = o_i)$ correspondant au comportement réflexe, le deuxième $P(O_A = o_i|S = percept(M))$ au comportement communicatif. Ce comportement mène à l'émergence d'un code de parole commun dans la société d'agents (avec un taux de compréhension de 100%) et possède les même propriétés que le comportement hybride : il réalise de la dispersion et exploite les non-linéarités présentes dans la fonction *percept*. Il nous semble cependant plus intéressant que le comportement communicatif car il introduit une contrainte articulatoire grâce au terme $P(M|O_L = o_i)$ et fournit une base mathématique à la théorie, ni perceptive ni motrice, mais perceptuo-motrice que nous défendons pour la communication parlée : la Théorie de la Perception pour le Contrôle de l'Action [10].

5. Conclusions et perspectives

Le cadre de simulation présenté ici montre comment on peut faire émerger des principes de mise en forme (tels que la théorie de la dispersion ou la théorie quantique) d'un principe de dérivation plus général, faisant dans notre cas un lien entre l'émergence du langage et un comportement plus primitif de déixis. Ce comportement est attesté et bien décrit chez les primates non humains [4, 11].

Nous avons commencé à implémenter ces principes sur un modèle réaliste de conduit vocal et de système auditif, les simulations conduisent à une dispersion au sein du triangle vocalique, conformément aux résultats classiques des travaux de Lindblom et suivants [6]. Pour passer de vocalisations simples comme les voyelles à des vocalisations plus complexes comme les syllabes, nous nous appuierons sur un autre comportement précurseur, celui des cycles mandibulaires associés à la mastication, et produisant, selon la théorie Frame/Content, un principe de modulation important pour passer à des séquences complexes [7].

Dans cette démarche, nous privilégions systématiquement la recherche de comportements précurseurs et de contraintes émergeant de ces comportements [10]. Nous nous distinguons par là d'autres travaux cherchant à faire émerger la complexité à partir d'un nombre d'hypothèses minimum [9]. Notre position est au contraire que les formes sont contraintes par les propriétés des systèmes en interface et que c'est en intégrant le plus grand nombre possible de connaissances sur ces systèmes (production, perception de vocalisations) que les simulations s'approcheront le plus des réalités linguistiques.

Références

[1] C. Abry, A. Vilain, and J.-L. Schwartz. Vocalize to localize ? a call for better crosstalk between auditory and visual communication systems researchers. *Interaction Studies : social behaviour and communication in biological and artificial systems*, 5(3) :313–325, 2004.

[2] A.-R. Berrah, H. Glotin, R. Laboissière, P. Bessière, and L.-J. Boë. From form to formation of phonetic structures : An evolutionary computing perspective. In *Int. Conf. on Machine Learning, Workshop on Evolutionary Computing & Machine Learning (ICML'96)*, 1996.

[3] C. Frith. *Neuropsychologie cognitive de la schizophrénie*. Paris, PUF, 1996.

[4] D.A Leavens and W.D. Hopkins. Deriving reference : bipedalism, encephalization, and the referential problem space. In *Vocoid (VOcalization, COmmunication, Imitation and Deixis in infant and adult human and non-human primates), Grenoble, May 14-16th.*, 2007.

[5] O. Lebeltel, P. Bessière, J. Diard, and E. Mazer. Bayesian robot programming. *Advanced Robotics*, 16 :49–79, 2004.

[6] J. Liljencrants and B. Lindblom. Numerical simulation of vowel quality systems : the role of perceptual contraste. *Language*, 48 :839–862, 1972.

[7] P. F. MacNeilage. The Frame/Content theory of evolution of speech production. *Behavioral and Brain Sciences*, 21 :499–546, 1998.

[8] C. Moulin-Frier. Jeux déictiques dans une société d'agents sensori-moteurs. Master's thesis, Grenoble-INP, 2007.

[9] P.Y. Oudeyer. Aux sources du langage : l'auto-organisation de la parole. *Cahiers Romans de Sciences Cognitives, In Cognito*, 2(2) :1–24, 2004.

[10] J.-L. Schwartz. Eléments pour une morphogenèse des unités du langage. In *Colloque Systèmes complexes en Sciences Humaines et Sociales, Cerisy (à paraître)*, 2008.

[11] K. Slocombe and K. Zuberbühler. Functionally referential communication in a chimpanzee. *Current Biology*, 15(19) :1779–1784, 2005.

[12] L. Steels. The synthetic modeling of language origins. *Evolution of Communication Journal*, 1(1) :1–34, 1997.

[13] K. Stevens. On the quantal nature of speech. *Journal of Phonetics*, 17(1/2) :4–45, 1989.